

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-066196

(43)Date of publication of application : 09.03.1999

(51)Int.Cl.

G06F 19/00

G06F 17/21

G06T 7/00

G06K 9/20

H04N 1/21

H04N 1/40

(21)Application number : 09-220426

(71)Applicant : RICOH CO LTD

(22)Date of filing : 15.08.1997

(72)Inventor : SAITO TAKASHI

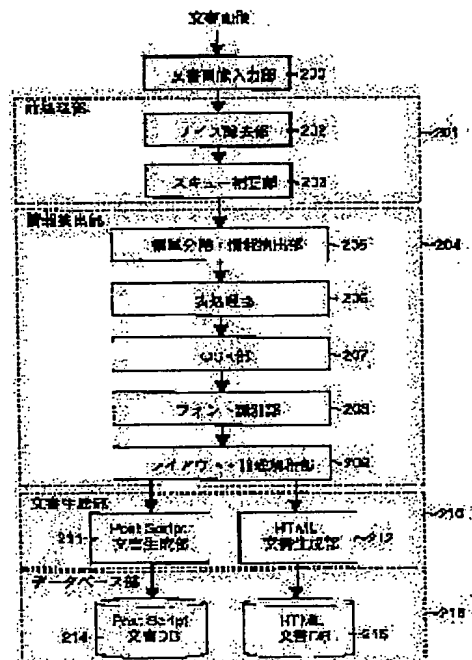
ABE TEI

KOUCHI TSUKASA

(54) DOCUMENT IMAGE RECOGNITION DEVICE AND COMPUTER-READABLE RECORDING MEDIUM WHERE PROGRAM ALLOWING COMPUTER TO FUNCTION AS SAME DEVICE IS RECORDED

(57)Abstract:

PROBLEM TO BE SOLVED: To generate documents in various format complying with usages such as a document whose reproduction is given priority and a document whose contents are made important.
SOLUTION: This device is equipped with a document image input part 200 which inputs a document image generated by optically reading a paper document, a preprocess part 201 which performs noise removal and a skew correcting process for the inputted document image, an information extraction part 204 which performs a recognizing and extracting process for a character area including character strings and/or a an image area including images of a graph, a table, a photograph or the like, and a character recognizing process for the character strings in the extracted character area and also analyzes the layout of the document image to extract layout information, a document generation part 210 which generates a PostScript document and an HTML document according to the character recognition result and layout information extraction result, and a data base part 213 which stores the generated PostScript document and HTML document respectively.



LEGAL STATUS

[Date of request for examination] 13.06.2002

[Date of sending the examiner's decision of rejection] 04.02.2003

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

特開平 11-66196

(43) 公開日 平成 11 年 (1999) 3 月 9 日

(51) Int. Cl. ⁸		識別記号		F 1	
G 0 6 F	19/00	G 0 6 F	15/22	G	
G 0 6 F	17/21	G 0 6 K	9/20	C	
G 0 6 T	7/00	H 0 4 N	1/21		
G 0 6 K	9/20	G 0 6 F	15/20	A	
H 0 4 N	1/21		15/70	Q	
審査請求 未請求・請求項の数 4		OL		(全 16 頁) 最終頁に続く	

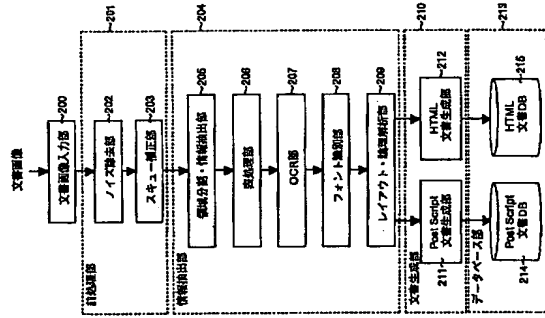
(21) 出願番号	特願平 9-220426	(71) 出願人	000006747 株式会社リコー
(22) 出願日	平成 9 年 (1997) 8 月 15 日	(72) 発明者	齋藤 高志 東京都大田区中馬込 1 丁目 3 番 6 号
		(73) 発明者	阿部 徳 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社 社リコー内
		(74) 発明者	幸地 司 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社 社リコー内
		(74) 代理人	弁理士 酒井 宏明 社リコー内

(54) 【発明の名称】 文書画像認識装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体

(57) 【要約】

【課題】 紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態の文書を作成すること。

【解決手段】 紙文書を光学的に読み取ることによって生成した文書画像を入力する文書画像入力部 200 と、入力した文書画像に対してノイズ除去およびスキュー補正処理を行う前処理部 201 と、文書画像から文字列を含む文字領域および/または図、表、写真等の画像を含む画像領域の認識・抽出処理、抽出した文字領域の文字列についての文字認識処理、および文書画像のレイアウトを解析し、レイアウト情報の抽出処理を行う情報抽出部 204 と、文字認識結果およびレイアウト情報抽出結果に基づいて、PostScript 文書および HTML 文書を作成する文書生成部 210 と、生成した PostScript 文書および HTML 文書をそれぞれ格納するデータベース部 213 と、を備えている。



【特許請求の範囲】

【請求項 1】 紙文書を光学的に読み取ることによって生成した文書画像を入力する入力手段と、

前記入力手段を介して入力した文書画像から文字列を含む文字領域および/または図、表、写真等の画像を含む画像領域を認識して抽出する領域抽出手段と、前記領域認識手段で抽出した文字領域の文字列について文字認識処理を行う文字認識手段と、

前記領域抽出手段の抽出結果に基づいて、前記文書画像のレイアウトを解析し、レイアウト情報を抽出するレイアウト情報抽出手段と、

前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、ページ記述言語を用いた第 1 の文書を作成する第 1 の文書生成手段と、

前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、構造化記述言語を用いた第 2 の文書を作成する第 2 の文書生成手段と、

前記第 1 および第 2 の文書生成手段で生成した第 1 および第 2 の文書をそれぞれ格納する格納手段と、

【請求項 2】 前記第 1 の文書は、PostScript 形式または PDF 形式によって表現された文書であることを特徴とする請求項 1 に記載の文書画像認識装置、

【請求項 3】 前記第 2 の文書は、SGML、HTML または XML によって表現された文書であることを特徴とする請求項 1 に記載の文書画像認識装置、

【請求項 4】 前記請求項 1 ~ 3 のいずれか 1 つに記載の文書画像認識装置の各手段としてコンピュータを機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体、

【発明の詳細な説明】

【0001】 発明の原する技術分野 本発明は、紙文書を光学的に読み取ることによって得た文書画像から文字コードを抽出するという単なる文字認識処理を行うのではなく、紙文書の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を作成することができるようにした文書画像認識装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【0002】 【従来の技術】 パーソナルコンピュータやネットワークの急速な普及は、オフィスにおける文書の電子化を促進することとなった。ところが、紙は閲覧性が良い等の利便性があるため、紙文書の生産は未だ止むことがなく、紙の形で保存・流通されている文書は未だ多量に存在し

ていることが現状であり、このギャップがオフィスワークにおける生産効率の低下の一因となっている。

【0003】 つまり、文書の作成、流通、閲覧、管理、再利用といった一連の流れにおいて、文書は紙に記録された形とコンピュータ上のデータという形で存在しており、このように文書が紙に記録された形とコンピュータ上のデータという形で存在しているのは、相互の変換コストが高いということが原因となっている。

【0004】 上記問題を解決する手段の一つとして、OCR (Optical Character Reader) を挙げることができる。この OCR を用いることにより、紙文書をスキャナ等で読み取って文書画像を生成した後、文書画像中の文字列を文字コードに変換することができる。

【0005】 【発明が解決しようとする課題】 しかしながら、上記 OCR は、文書画像中の文字列を文字コード情報に変換することは可能であるが、元の紙文書のレイアウト等まで抽出することができないため、OCR によって生成された文書は単なるテキストの題列であって、元の文書画像が持つ様々な情報を十分に利用することができないという問題点があった。換言すれば、文書は、作成者の意図に応じて様々なレイアウト処理が施されているが、単に OCR を用いただけでは、文書のレイアウトを抽出することができず、元の文書のレイアウトを再現した文書が新たに生成したり、元の文書のレイアウトを利用して新たなレイアウトの文書を作成したりすることはできなかった。

【0006】 本発明は上記に鑑みてなされたものである。本発明は、紙文書を光学的に読み取ることによって得た文書画像から文字コードを抽出するという単なる文字認識処理を行うのではなく、紙文書の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を作成することができるようにすることを目的とする。

【0007】 【課題を解決するための手段】 上記目的を達成するため、請求項 1 の文書画像認識装置は、紙文書を光学的に読み取ることによって生成した文書画像を入力する入力手段と、前記入力手段を介して入力した文書画像から文字列を含む文字領域および/または図、表、写真等の画像を含む画像領域を認識して抽出する領域抽出手段と、前記領域認識手段で抽出した文字領域の文字列について文字認識処理を行う文字認識手段と、前記領域抽出手段の抽出結果に基づいて、前記文書画像のレイアウトを解析し、レイアウト情報を抽出するレイアウト情報抽出手段と、前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、ページ記述言語を用いた第 1 の文書を作成する第 1 の文書生成手段と、

前記文字認識手段による文字認識結果およびレイアウト情報抽出手段によるレイアウト情報抽出結果に基づいて、構造化記述言語を用いた第 2 の文書を作成する第 2 の文書生成手段と、

前記第 1 および第 2 の文書生成手段で生成した第 1 および第 2 の文書をそれぞれ格納する格納手段と、

【請求項 2】 前記第 1 の文書は、PostScript 形式または PDF 形式によって表現された文書であることを特徴とする請求項 1 に記載の文書画像認識装置、

【請求項 3】 前記第 2 の文書は、SGML、HTML または XML によって表現された文書であることを特徴とする請求項 1 に記載の文書画像認識装置、

【請求項 4】 前記請求項 1 ~ 3 のいずれか 1 つに記載の文書画像認識装置の各手段としてコンピュータを機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体、

白部分) および縦線を検出し、検出した距離および縦線の本数と共に、領域分割処理で分割した文字領域の位置に関する属性情報に基づいて、段組種類の判定を行う。
【0030】その後、検出処理部206は、領域分割・情報抽出部205で分割した各領域の属性情報に基づいて、表を含む領域が存在するか否かを判定する(S306)。この判定は、上記のように装置側で自動的に行うことにしても良いし、ユーザが指定しても良い。表が含まれていない場合には、ステップS308に進む。
【0031】一方、表が含まれている場合、検出処理部206は、表の枠と縦線の構造を抽出すると共に、枠内の文字領域を抽出する(S307)。このように、表の中から文字領域を抽出することにより、次のOCR部207において、表中の文字認識処理を行うことができる。
【0032】そして、OCR部207は、領域分割・情報抽出部205および後処理部206で分割した文字領域の文字列について文字認識処理を行う(S308)。すなわち、OCR部207は、文字領域について、行切り出しおよび文字切り出し処理を行い、文字切り出した個々の文字パターンについて文字認識処理を行う。加えて、OCR部207は、文字認識結果に対して、言語処理による誤り補正を行う。
【0033】フォント識別部208は、領域分割・情報抽出部205で分割した文字領域の文字列について、行単位でフォントが強調系(印ゴシック等) や非強調系(印刷体等) のいずれであるかを識別する処理を行う(S309)。具体的には、例えば、黒画密度やランレングスの分布等に基づいて、フォントの特徴を識別する。
【0034】続いて、レイアウト・輪理解析部209は、上記各部による処理の結果に基づいて、文書画像のレイアウトを解析する(S310)。ここで行われるレイアウトの解析処理には、例えば、タイトル部、小見出し部、キャプション、ヘッダ・フッタ部の検出処理が含まれる。

【0035】ここで、タイトルは、一般的に本文の文字とはサイズや行ピッチが異なり、また、存在する位置も本文とは若干離れていることから、領域分割・情報抽出部205で付与した領域の位置に関する属性情報や、フォント識別部208による識別結果を用いて、タイトル部を検出することができる。
【0036】小見出しは、本文の文字と文字サイズがほぼ等しい場合も多く、本文に近接した場所に位置することから、本文と同一の領域に存在していることも多い。そこで、各文字領域の先行行の文字サイズまたはフォントが、同一の文字領域中の他の文字のものと異なる場合に、先行行を小見出しと判定する。

【0037】また、キャプションは、図、表、写真等の画像に付与されたものであり、一般的に画像領域の近傍で、本文とは離れた位置に存在すること、さらには、

【0044】図5は、PostScript文書を画面表示した様子の一例を示す説明図である。図5に示すように、図4に示した文書画像に基づいてPostScriptで表現した文書を生成することにより、元の紙文書と同一のレイアウトの文書を容易に生成することができ、すなわち、紙文書を再現して文書を生成することにより、文書作成者の意図をできるだけ正確に伝えることができるような、流通および閲覧に適した文書を得ることができる。なお、PostScript文書を画面表示した様子は元の紙文書とほぼ同一であるが、内部情報は保持されているため、検索を行ったり、再利用したりすることができ、

【0045】また、図6は、HTML文書を画面表示した様子の一例を示す説明図である。図6は、図と図番号のハイパーテキスト化を行ったものであり、例えば、本文中の「図9」をマウス等でクリックすると、「図9」に該当する図が画面表示される。このように、紙文書が持つレイアウトにとらわれないことなく、ハイパーテキスト化するることにより、紙文書の内容を優先したコンピュータならではの形を持った文書を生成することができ、

【0046】このように、実施の形態1の文書画像認識装置によれば、文書画像から文字コードを抽出するといふ単なる文字認識処理を行うだけではなく、文書画像の持つ様々な情報を抽出して利用するため、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を生成することができる。

【0047】なお、図1においては、ネットワーク109を介したシステムとして実施の形態1の文書画像認識装置の構成を説明したが、図2に示す機能を1台のコンピュータで持たせることにより、スタンドアローンの形態で文書画像認識装置を構成することもできる。

【0048】また、実施の形態1の文書画像認識装置では、上述したPostScript文書やHTML文書を生成することにしたが、これらに限定するものではなく、必要に応じて他の形式の文書を生成することもできる。

【0049】(実施の形態2) 実施の形態2の文書画像認識装置は、実施の形態1で説明したようにして生成したPostScript文書やHTML文書を効率よく検索することができるようにしたものである。具体的に、OCR部207で文字認識した文字列から所定の文字列を抽出し、抽出した文字列を対応するPostScript文書やHTML文書に関連づけおき、該当する文字列を検索することにより、関連づけられたPostScript文書やHTML文書を検索結果として出力力できるようにするものである。以下では、この文字列のことをキーテキストと定義することにする。

【0050】図7は、実施の形態2の文書画像認識装置

の概念構成図である。図7において、実施の形態1で説明した図2と同一の構成については同一の符号を付すことにし、それらの詳細な説明については省略する。

【0051】実施の形態2の文書画像認識装置は、図7に示すように、OCR部207で文字認識した文字列から上述したキーテキストを抽出するキーテキスト抽出部700と、キーテキスト抽出部700で抽出したキーテキストを入力し、キーテキスト登録部701と、検索要求を入力し、キーテキストDB702に登録されたキーテキストを検索し、該当するキーテキストに関連づけられたPostScript文書またはHTML文書を検索結果として出力する検索処理部703と、を備えている。

【0052】なお、PostScript文書DB214、HTML文書DB215およびキーテキストDB702は、検索処理部703に設けられる。この検索処理部703は、図1における検索サーバ108に該当し、クライアント107からの検索要求に基づいて、上記検索処理を行う。また、実施の形態2の文書画像認識装置をスタンドアローンの形態で構成した場合には、直接検索要求を入力して検索処理を行う。

【0053】上記キーテキスト抽出部707で抽出するキーテキストとしては、文書を論理的に表した文字列、例えば、文書全体、章、節のタイトルや、ヘッダ・フッタ等の思想的事項、文書の要約文等が考えられる。また、文書中の図等を基調として、図のキャプションを構成する文字列や、図等を含むセンテンス、このセンテンスを含むパラグラフおよびページ単位の文字列をキーテキストとして抽出しても良い。なお、上記キーテキストを抽出するに際しては、文書画像のレイアウトを解析する必要があることから、キーテキスト抽出部700は、レイアウト・輪理解析部209による解析結果を用いて、キーテキストの抽出処理を行うようにしても良い。

【0054】また、キーテキスト抽出部701は、上記キーテキスト抽出部700で抽出したキーテキストを入力し、入力したキーテキストを文書生成部210で生成したPostScript文書やHTML文書に関連づけ、キーテキストDB702に格納する。

【0055】さらに、検索処理部703は、検索要求を入力すると共に、検索結果を出力する入出力部704と、入出力部704から検索要求を入力し、キーテキストDB702から該当するキーテキストを検索する検索エンジン705とを有している。具体的に、入出力部704は、検索要求を入力して検索エンジン705に検索要求を出力する。検索エンジン705は、入出力部704から検索要求を入力し、キーテキストDB702から該当するキーテキストを検索し、該当するキーテキストを入出力部704に出力する。入出力部704は、入力したキーテキストに関連づけられたPostScript文書やHTML文書を検索結果として出力する。

【0056】このように、実施の形態2の文書画像認識装置100は、文書画像中の文字列をキーテキストとして文書の検索を行うことにしたため、検索要求に対して、最も適切な検索結果を得ることができると共に、所望な検索処理を実現することができる。

【0057】なお、上述した実施の形態2においては、キーテキストを用いて文書の校閲を用いることにしたことが、キーテキストに代え、OCR部207で文字認識した結果全てを用いて全文校閲を行うようにすることもできる。図8は、実施の形態2の文書画像認識装置の外形例を示す概念構成図である。

【0058】図8に示すように、上述したキーテキスト抽出部700、キーテキスト登録部701およびキーテキストDB702に代えて、テキスト登録部800およびテキスト検索用テキストDB801を設け、テキスト登録部800がOCR部207から文字認識結果、即ち文書画像中のテキストの全文を入力、入力したテキストを文書生成部210で生成したPosiScript文書やHTML文書に関連づけ、全文検索用テキストDB801に登録し、HTML文書を用いて、PosiScript文書801に登録されたテキストを用いて、PosiScript文書やHTML文書を検索することができる。この場合は、テキストを全文検索用テキストDB801に登録するため、検索の便に各ファイルのオープン・クロージングという処理が必要となるため、検索処理の効率化を図ることができ、

【0059】（実施の形態3）続いて、実施の形態3の文書画像認識装置について説明する。実施の形態3の文書画像認識装置は、実施の形態1のものと同様、流通経路におよび閲覧者、利用者を考慮した文書を生成するものである。

【0060】実施の形態3の文書画像認識装置における流通および閲覧を考慮した文書とは、オリジナルの紙式文書を読み取って生成した文書画像であり、ここではデジタルデータと定義することによる。また、利用を考慮した文書とは、文書画像中の文字列について文字認識を行った結果であり、ここではテキストデータと定義することによる。

【0061】図9は、実施の形態3の文書画像認識装置1の概念構成図である。なお、図9において、実施の形態1で説明した構成と同一の構成については同一の符号を付し、これらの詳細な説明は省略する。

【0062】実施の形態3の文書画像認識装置は、大
く文書画像入力部200、前処理部201、情報抽出部202、登録部203およびデータベース部213から構成される。なお、文書画像入力部200、前処理部201、情報抽出部202および登録部203は、図1に示した文書画像処理サーバ105に該当し、データベース部213は、図1に示した検索サーバ108に該当する。

【0063】文書画像入力部200および前面部201については実施の形態1で説明した通りである。また、情報抽出部204においては、PostScript文書やHTML文書を生じないため、フォント識別部208およびクライアント・論理解析部209が用意されている以外、実施の形態1で説明した通りである。

【0064】実施の形態1の文書画像認識装置と異なるのは、イメージD3登録部901およびテキストD8登録部902を備えた登録部900と、イメージD8904およびテキストD8905を備えたデータベース部903である。

【0065】イメージDB登録部901は、前処理部901からノイズ除去およびスキュー補正を行った後のイメージデータを入力、入れたイメージデータをデータベース部903のイメージDB904に登録する処理を行い、テキストDB登録部902は、OCR部207で文字認識処理を行った結果であるテキストデータをデータベース部903のテキストDB905に登録する処理を行う。なお、同一タイプのイメージデータとテキストデータの両方を登録する場合には、一方から他方を呼び出すことができるように、相互に関連づけを行うことにも良い。

【0066】なお、イメージDB9004およびテキストDB9005にそれぞれ格納されたイメージデータおよびテキストデータは、図面表示を行うことによつて各クライアント107で図面表示することができ、すなわち、クライアント107からの検索要求に応じて、検索サーバ108がイメージDB9004およびテキストDB9005から該当する文書を検索して出力し、クライアント107は、検索サーバ108から検索結果を入力し、検索要求に該当する文書を画面に表示する。

【0067】このように、実施の形態3の文書画像認識装置によれば、イメージデータをデータベースに登録するため、紙文書そのものをコンピュータ上で画面表示することができ、また、テキストデータをデータベースに登録するため、紙文書中の情報と容易に利用することができる。

【0068】（実施の形態4）実施の形態4の文書画像認識装置は、実施の形態3で説明したようにして生成したイメージデータやテキストデータを、実施の形態2で説明したキーテキストを用いて効率的に検索することができようとしたものである。

【0069】図10は、実施の形態4の文書画像認識装置の概念構成図である。図10において、実施の形態1で説明した図2および実施の形態2で説明した図7と同様の構成については同一の符号を付すことにし、それらの詳細な説明については省略する。

【0070】実施の形態4の文書画像認識装置は、図10に示すように、文書画像のレイアウトを解析し、OCR部207で文字認識した文字列から上述したキーテーキ

キータを抽出するキータキスト抽出部700と、キータキスト抽出部700で抽出したキータキストを入力し、キータキストDB702に登録するキータキスト登録部701と、直接に検索要求を入力し、キータキストDB702に登録されたキータキストを検索して、該当するキータキストに照應づけられたメタデータやキータキストデータを検索結果として出力する検索処理部703と、
を備える。

【0071】上記キーテキスト抽出部207で抽出するキーテキストとしては、文書を断片的に表した文字列、例えば、文書全体、章、節のタイトルや、ヘッダ・フッタ等の母題的事項、文書の要約文等が与えられる。また、文書中の図等を基準として、図のキャプションを構成する文字列や、図番を含むセンテンス、このセンテンスを含むパラグラフおよびページ単位の文字列をキーテキストとして抽出しても良い。

【0072】また、キーデキスト抽出部701は、上記キーデキスト抽出部700で抽出したキーデキストを入力し、対応するイメーজデータやキーデキストデータに関連づけ、キーデキストDB702に格納する。

[0073] さらに、検索処理部703は、検索要求を入力すると共に、検索結果を出力する入力部704と、入力部704から検索要求を入力し、キーテキストDB702から該当するキーテキストを検索する検索エンジン705とを有している。具体的に、入力部704では、検索要求を入力して検索エンジン705に検索エンジンを介して検索要求を入力して検索エンジン705は、入力部704から検索要求を入力し、キーテキストDB702から該当するキーテキストを検索し、該当するキーテキストの入力部704に出力する。入力部704は、入力をしたキーテキストに関連づけられたイメーデータやタグデータを検索結果として出力する。

キ

【0074】上記検索処理部703は、図1における検索サーバ108に該当し、クライアント107からの検索要求に基づいて、検索処理を行う。また、実施の形態2の文書画像認識装置をスタンドアローンの形態で構成した場合には、直接検索要求を入力して検索処理を行う。

【0075】このように、実施の形態4の文書画像認識処理装置100は、文書画像中の文字列をキーデキストとして表照によれば、文書画像中の文字列をキーデキストとして文書の検索を行うことにしたため、検索要求に対して、最も適切な検索結果を得ることができると共に、高度な検索処理を重視することができ、

【0076】なお、上述した実施の形態4においては、キーキストを用いて文書の検索を用いることにした。キーキストに代え、OCR部207で文字認識した結果全てを用いて全文検索を行うようにすることもできる。図11は、実施の形態4の文書画像認識装置の形例を示す概念構成図である。

【0077】図11に示すように、テキスト登録部8

文書画像中のテキストデータを全文を入力し、対応するイメージデータやテキストデータに関連づけし、全文検索用テキストDB8001に登録する。その結果、全文検索用テキストDB8001に登録されたテキストを用いて、イメージデータやテキストデータを検索することができ、テキストDB8001に登録されたテキストを検索することによって、テキストを全文検索用テキストDB8001に登録するため、検索の度に各ファイルのオープン・クローズという処理が不要となるため、検索処理の高速化を図ることができ、

像図表置換は、それぞれ任意に組み合わせて1つの文書画像置換部を構成することができ、

【0079】さらに、上述した実施の形態1～4の文部画像認識装置としてコンピュータを機能させるプログラムを作成し、これをハードディスク、フロッピーディスク、CD-ROM、MO、DVD等のコンピュータ読み取り可能な記録媒体に記録して、記録媒体を介してプログラムを配布することができる。そして、記録媒体に記録されたプログラムをコンピュータで読み出して実行することにより、上述した文部画像認識装置を実現することができる。

【発明の効果】以上説明したように、本発明の文書画像認識装置（請求項1）によれば、紙文書を光学的に読み取り、入力手段として入力した文書画像を入力する入力手段と、入力手段が生じた文書画像から文字列を含む文字領域および／または図表、写真等の画像を含む画像領域を認識して抽出する領域抽出手段と、領域認識

手段で抽出した文字領域の文字列について文字認識処理を行う文字認識手段と、領域抽出手段の抽出結果に基づいて、文書画像のレイアウト情報と、レイアウト情報とを抽出するレイアウト情報抽出手段と、抽出したレイアウト情報に基づいて、ページ記述言語を用いた第1の文書と生成する第1の文書生成手段と、文字認識手段による文字認識結果およびレイアウト情報抽出手段による文字認識結果とレイアウト情報抽出手段とを用いた第2の文書と生成する第2の文書生成手段と、第1および第2の文書と生成手段で生成した第1の文書と第2の文書とをそれぞれ格納する格納手段と、を備えたため、文書画像から文字コードを抽出するという単なる文字認識処理を行うだけではなく、文書画像の待機した様々な情報を抽出して利用することにより、紙文書等、再現を要した様式や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書と生成することができ、

【0081】また、本発明の文書画像認識装置（請求項2）によれば、請求項1に記載の文書画像認識装置にお

いて、第1の文書は、PostScript形式またはPDF形式によって表現された文書であるため、文書作成者の意図をできるだけ正確に伝えることができる。な、流通および閲覧に適した文書を得ることができる。

【0082】また、本発明の文書画像認識装置（請求項3）によれば、請求項1に記載の文書画像認識装置において、第2の文書は、SGML、HTMLまたはXMLによって表現された文書であるため、紙文書が持つレイアウトにとらわれないこと、紙文書の内容を優先したコンピュータならではの形を持った文書を作成することができる。

【0083】さらに、本発明のコンピュータ読み取り可能な記録媒体（請求項4）によれば、請求項1〜3のいずれか一つに記載の文書画像認識装置の各手段としてコンピュータを機能させるためのプログラムを記録したため、記録したプログラムをコンピュータ上で実行することにより、文書画像の持つ様々な情報を抽出して利用することにより、紙文書の再現を優先した文書や紙文書の内容を重視した文書等、利用目的に応じた様々な形態のコンピュータ上で利用可能な文書を作成することができる。

【図面の簡単な説明】

- 【図1】実施の形態1の文書画像認識装置のシステム構成を示す構成図である。
- 【図2】図1に示す文書画像認識装置の概念構成図である。
- 【図3】実施の形態1の文書画像認識装置の動作手順を示すフローチャートである。
- 【図4】実施の形態1の文書画像認識装置において、文書画像入力部を介して入力した文書画像を画面表示した様子の一例を示す説明図である。
- 【図5】実施の形態1の文書画像認識装置において、生成したPostScript文書を画面表示した様子の一例を示す説明図である。
- 【図6】実施の形態1の文書画像認識装置において、生成したHTML文書を画面表示した様子の一例を示す説明図である。
- 【図7】実施の形態2の文書画像認識装置の概念構成図である。
- 【図8】実施の形態2の文書画像認識装置の変形例を示す概念構成図である。
- 【図9】実施の形態3の文書画像認識装置の概念構成図である。

である。

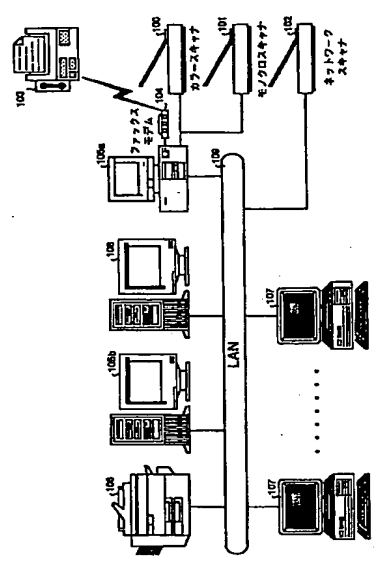
【図10】実施の形態4の文書画像認識装置の概念構成図である。

【図11】実施の形態4の文書画像認識装置の変形例を示す概念構成図である。

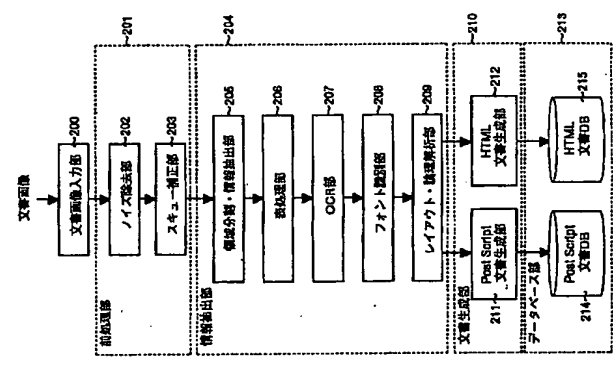
【符号の説明】

- 100 カラースキャナ
- 101 モノクロスキャナ
- 102 ネットワークスキャナ
- 103 ファクシミリ装置
- 104 ファックスモデム
- 105 (105a, 105b) 文書画像処理サーバ
- 106 デジタル複合機
- 107 クライアント
- 108 検索サーバ108
- 109 ネットワーク
- 200 文書画像入力部
- 201 前処理部
- 202 ノイズ除去部
- 203 スキュー補正部
- 204 情報抽出部
- 205 傾斜分割・情報抽出部
- 206 表処理部
- 207 OCR部
- 208 フォント識別部
- 209 レイアウト・輪理解析部
- 210 文書生成部
- 211 PostScript文書生成部
- 212 HTML文書生成部
- 213 データベース部
- 214 PostScript文書DB
- 215 HTML文書DB
- 700 キーキスト抽出部
- 701 キーキスト登録部
- 702 キーキストDB
- 703 検索処理部
- 704 入力部
- 705 検索エンジン
- 800 テキスト登録部
- 801 全文検索用テキストDB

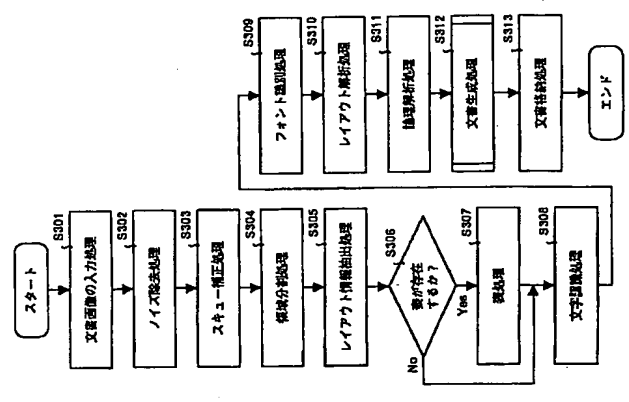
【図1】



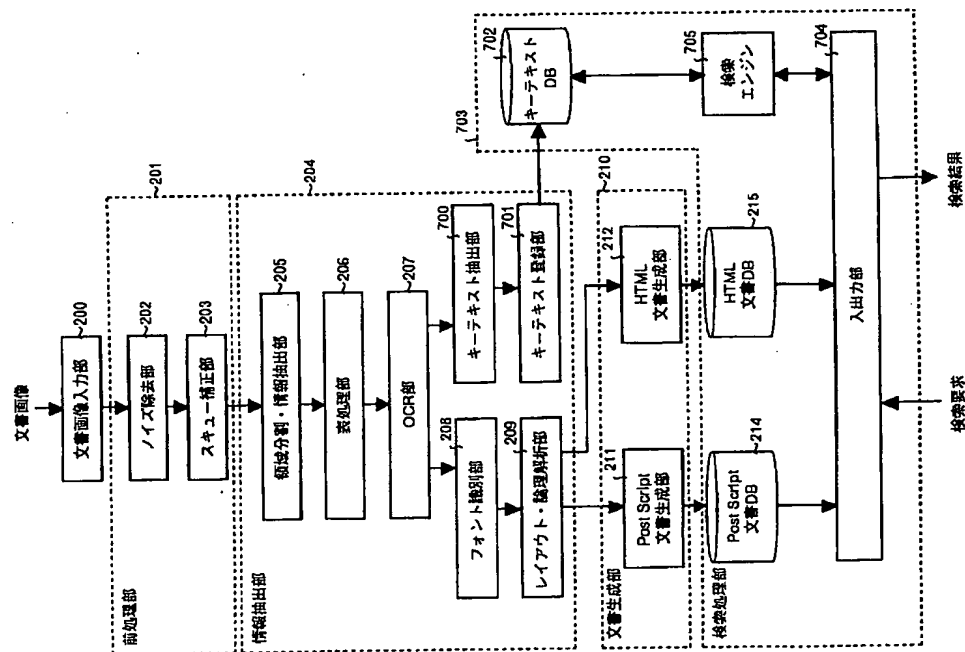
【図2】



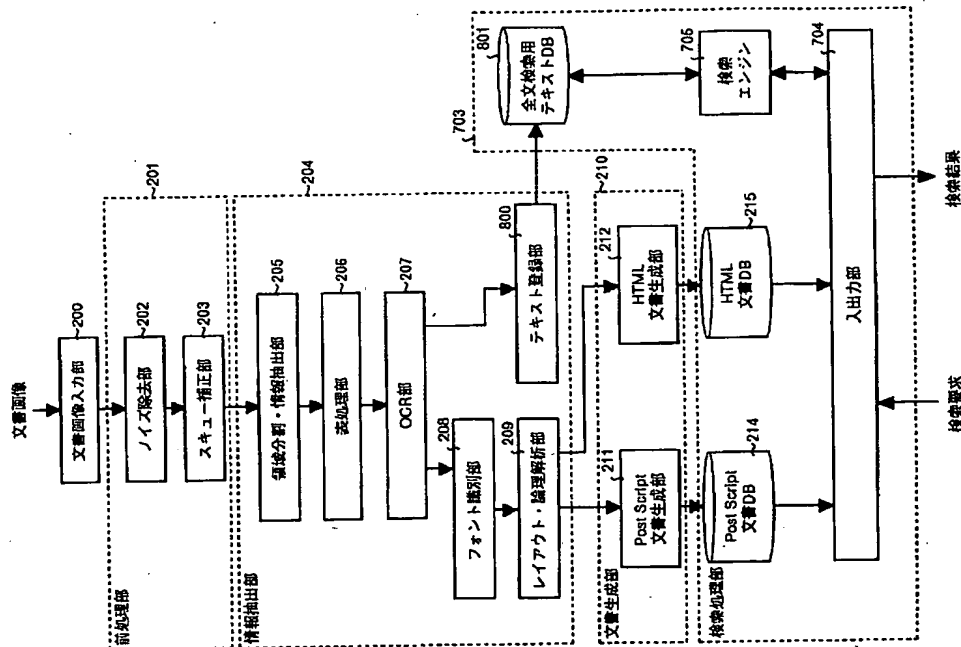
【図3】



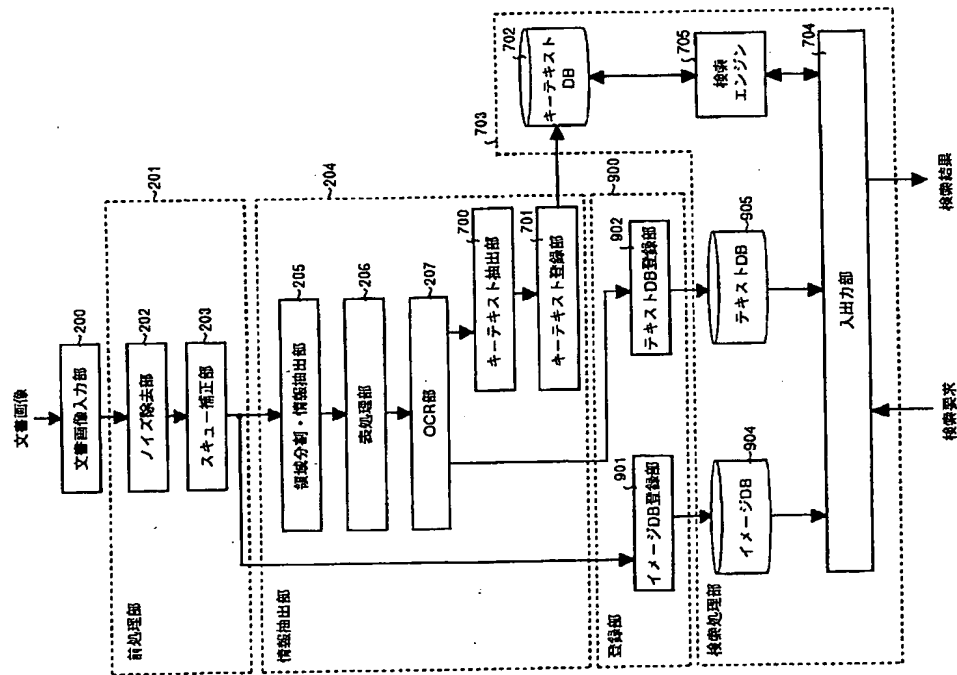
【図7】



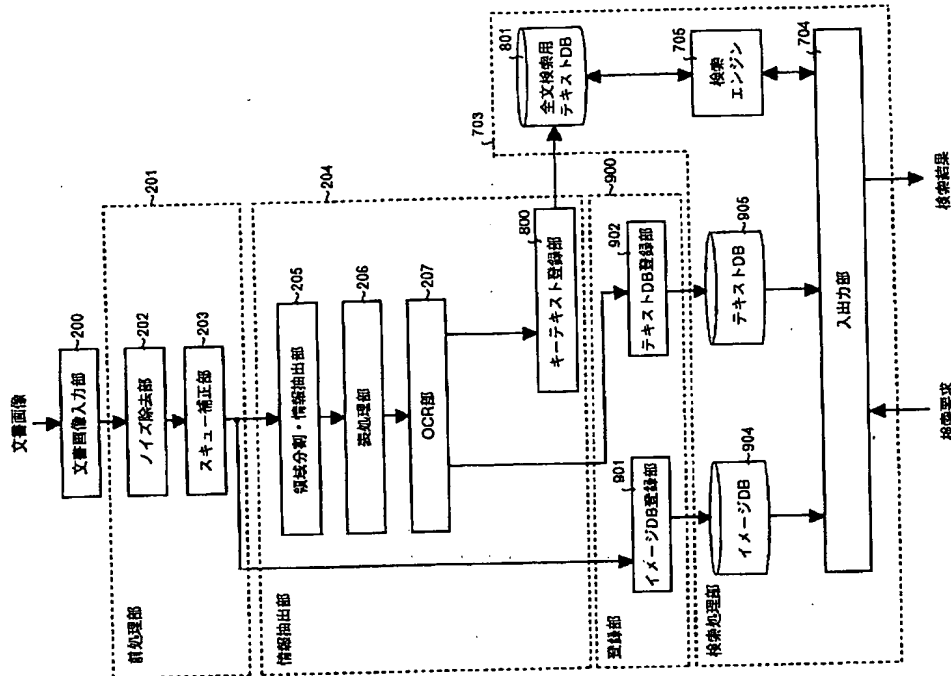
【図8】



【図10】



【図11】



フロントページの続き

(51) Int. Cl. °

H04N 1/40

識別記号

F I H04N 1/40 F